Associations between Body Mass Index (BMI) and Accelerometer Time Series Data: National Health and Nutritional Examination Survey (NHANES) 2005-2006

by
Luchao Qi

A thesis submitted to The Johns Hopkins University in conformity
with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland
May, 2020

Abstract

Higher rates of impact physical activity (PA) are thought to promote weight loss and health. However, little cross-sectional epidemiological research exists relating average activity and weight, obesity or body mass index (BMI). Wearable technology provides a reliable measure of the frequency, duration, intensity, and timing of PA. The objective of this study was to examine the combined association between body mass index (BMI) and PA in a large cross-sectional study. The National Health and Nutrition Examination Survey (NHANES) has collected PA and demographic data of American participants in 2005. Linear regression was used to explore the association between BMI and accelerometry data with multiple covariates including age, race, and sex. In addition, a random forest model to predict BMI was implemented and the variable importance of PA was assessed relative to other predictors. Our results showed that BMI was associated with PA, in addition to race and age. Random forest variable importance metrics suggested age as having the greatest impact on predicting BMI followed by five principal component directions of variation of daily PA. In summary, the results show a consistent and strong relationship between physical activity and body mass index.

Thesis Readers

Dr. Brian S. Caffo (Primary Advisor)
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Dr. Joshua T. Vogelstein
Professor
Department of Biomedical Engineering
Johns Hopkins University

Dr. Jiawei Bai
Assistant Scientist
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Acknowledgements

This work was supported by grants from the National Institutes of Health (2R01NS060910-04A1) and the National Institutes of Health (5R01EB012547-02).

We would like to thank the Centers for Disease Control and Prevention (CDC) and the National Center for Health Statistics (NCHS) for permission to repost the publicly available NHANES and NDI data in the analytic format. We would also like to thank Mr. Andrew Leroux for collecting and organizing this data.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Obesity continues to be a global public health concern. Over the past two decades, obesity rates have increased worldwide and remained the highest in the United States [1]. Body mass index (BMI; kg/m2) is commonly used to classify overweight (BMI 25.0-29.9), obesity (BMI greater than or equal to 30.0) and extreme obesity (BMI greater than or equal to 40) among adults (age 20 years and over). According to National Health and Nutrition Examination Survey (NHANES) 2005-2006, an estimated 32.7 percent of U.S. adults 20 years and older are overweight, 34.3 percent are obese and 5.9 percent are extremely obese [2]. Physical activity is a major determinant of human health [3] and obesity has been shown to be inversely associated with physical activity [4]. However, the association between obesity and accelerometry based physical activity has not been widely examined in an ethnically diverse sample of adults in the United States [4]. Given the high prevalence of overweight and obesity among Americans [5], a better understanding of the relationship between BMI and physical activity would help clarify its influence on health and weight.

Unfortunately, physical activity patterns are difficult to study outside of laboratory settings. Self-reported interview data are often of limited value because respondents' perceptions of activity intensity vary and periods of physical activity are difficult for respondents to recall and quantify [6]. To address this problem, NHANES added the physical activity monitor (PAM) component and lifestyle patterns for 10348 subjects in 2005. The purpose of this study was to examine the associations between PA and BMI in a sample of Americans using regression models.

Chapter 2

Methods

Data

In NHANES, physical activity was measured using hip-worn accelerometers [7], which allows the study of the association between accelerometer-based objective PA measurements and BMI. The NHANES 2005–2006 data can be downloaded, processed, and combined with the survey Body Mass Index (BMI) using R package rnhanesdata[8, 9].

Using hip-worn accelerometers, NHANES collected the magnitude of acceleration or intensity of objective movement in the 2005–2006 wave. The accelerometry data consisted of sequential minute by minute records of activity intensity beginning from the time the device was initialized. The monitors were programmed to record minute-level resolution of accelerometry data beginning at 12:01 a.m. the day after health examination. Walking and similar types of activity are thought to be the main source of physical activity for most individuals [10]. NHANES provides the following activity measurements for each participant (1) PAXCAL: device calibration flag; (2) PAXSTAT: data reliability status flag; (3) WEEKDAY: day of the week; (4) MIN1-MIN1440: activity count corresponding to each minute of the day. For example, MIN1 is the activity count for 00:00-00:01. Figure 1 shows a scheme of the processed activity count matrix [11].

Based on data quality indicators (PAXCAL, PAXSTAT), wear/non-wear flags are

	SEQN	PAXCAL	PAXSTAT	WEEKDAY	MIN1	MIN2	⋯	MIN1440
	31128	1	1	1	166	27	⋯	0
	31128	1	1	2	0	0	⋯	0
(a)	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	31128	1	1	7	0	0	⋯	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮
	31193	2	1	2	0	0	⋯	1921
(b)	31193	2	1	3	335	2598	⋯	46
	31193	2	1	4	0	0	⋯	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(c)	31880	2	2	2	32767	32767	⋯	32767
	31880	2	2	3	32767	32767	⋯	32767
	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮
(d)	32008	1	2	5	0	0	⋯	0
	32008	1	2	6	NA	NA	⋯	NA

Figure 1. Scheme of the processed 2005-2006 wave accelerometry data. (a) corresponds to a subject with 7 full days of activity data, no data quality flag(PAXCAL, PAXSTAT) indicators, and no missing data; (b) shows a subject where the device was marked as uncalibrated upon return to NHANES; (c) presents a subject with both an uncalibrated device and data reliability issue; and (d) illustrates a subject with their last two days of data missing (other missing days not shown) [11].

estimated corresponding to each minute of the day. The format of the flag matrix is the same as the activity count matrix, except instead of reporting activity counts in the MIN1-MIN1440 columns, it provides a wear/non-wear flag indicator. These wear/non-wear flags were calculated by applying the classification algorithm [12]. These flags can take on the following 3 values: (1) 0: indicating that a particular minute is determined to be "non-wear"; (2) 1: indicating that a particular minute is determined to be "wear"; (3) NA: indicating that particular minute was missing data in the activity count data matrix used to create this set of wear/non-wear flags. NHANES also collected body measurement data (e.g. survey weights, ages, BMIs) to examine the associations between body weight and the health and nutritional status of the United States population. Those demographic and lifestyle data such as age, sex(Male, Female), ethnicity (Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Other) and BMI can be extracted according to unique participant identifier (SEQN). In this analysis, we would use the demographics data file of participants aged between 6 and 85 years old.

Manipulation

Criteria

To conduct the analysis in this manuscript, we excluded individuals who were (1) wearing the device fewer than 600 minutes (10 hours) each day; (2) not recorded as "Data deemed reliable" in the data reliability indicator variable (PAXSTAT); (3) not recorded as "The device was calibrated when it was returned by the participant" in the accelerometer calibration indicator variable (PAXCAL); (4) not wearing the device for 7 consecutive days [13]. This set of exclusion criteria yielded a sample size of 1,985 participants. After applying the exclusion criteria, there were some minutes identified as non-wear, but having non-zero activity counts. We component-wise multiplied the flag matrix (0 for "non-wear", 1 for "wear") with the activity count matrix so any periods identified as "non-wear" are replaced with 0 activity counts while other "wear" periods remain unchanged. We assume that the majority of non-wear time corresponds to either sleep or sedentary behaviors, though this assumption is untestable in practice [11].

Average daily activity pattern

For each participant, the activity counts across 7 days were aggregated by taking their average at each minute, respectively. This transformation ignores the potential effect of day-to-day and day of week variability on PA. The transformed matrix contains daily PA for each participant. That is, for N participants with 1,440 (60 * 24) minutes accelerometry data, the accelerometry data is originally stored as a (7N) $\times$1,440 data matrix, where 7 refers to the number of days each subject was instructed to wear the accelerometer device and 1,440 corresponds to the number of minutes in a day. After transformation, the accelerometry data will be stored as a (1*N) $\times$1,440 data matrix.

Model

Feature Construction

A linear relationship is hypothesized between BMI and a group of predictor variables [14]. Age/sex and race/ethnicity are complex traits that are particularly relevant for adjustment because each includes the social dimensions necessary for understanding the impact of physical activity on health [15]. Previous studies have shown BMI is not independent of age and sex and also a race-related difference among women [16]. More specifically, these covariates may substantially affect BMI, and PA and thus will be considered as potential confounders necessary for adjustment in regression models. To reduce the dimensionality of the PA data, we conducted principal component analysis (PCA) on the daily PA matrix. In Figure 2, we transformed correlated variables in the daily PA matrix into a smaller number of uncorrelated principal components (PCs). The first principal component is the scaled linear combination of the activity counts that accounts for the greatest amount of observed PA variability. Each succeeding component accounts for as much of the remaining variability as possible with scaled linear combinations mutually orthogonal to the previous. The transformed activity data (scores) after PCA, along with covariates, were used in multiple linear regression models to predict BMI.

To investigate the associations between variables, we used correlation analysis between our response and predictors. The correlation coefficient indicates a measurement of the direction and strength of a relationship between two variables linearly adjusted for the remainder. Figure 3(a) shows that for BMI, age and PC1 are the two most correlated variables with correlation coefficients, 0.41 and -0.39, respectively. In addition, a random forest machine learning algorithm was implemented to estimate feature importance when predicting BMI. In Figure 3 (b), the random forest results indicate that across all of the trees considered in the random forest, age and PC1 are

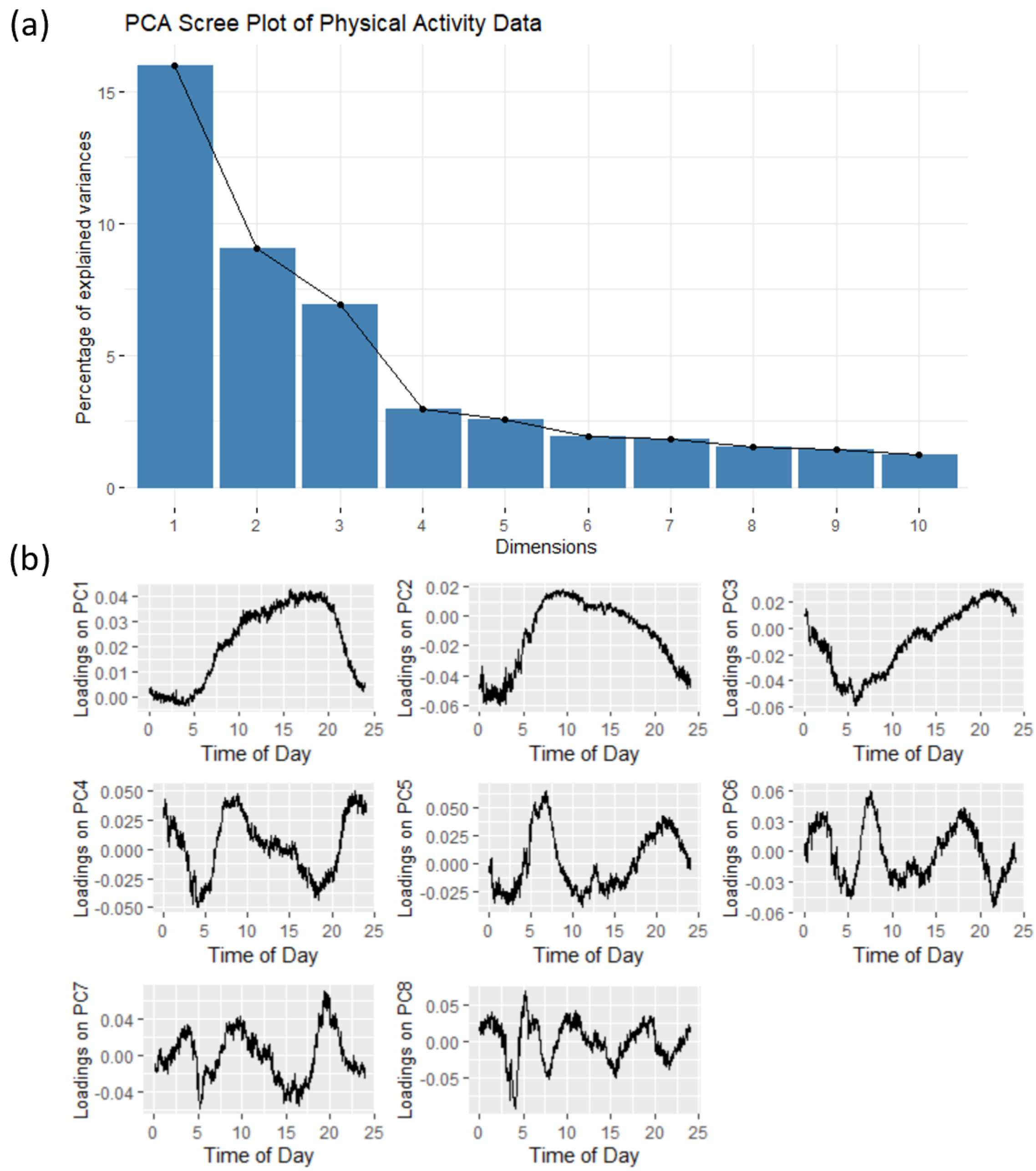

Figure 2. (a) Scree plot displaying the percentage of variation explained by each principal component. (b) First 8 principal components calculated on the minute-level NHANES accelerometry data. Solid lines represent principal component loadings of different hours.

(by far) the two most important variables, which agrees with our correlation analysis results.

(a)

(b)

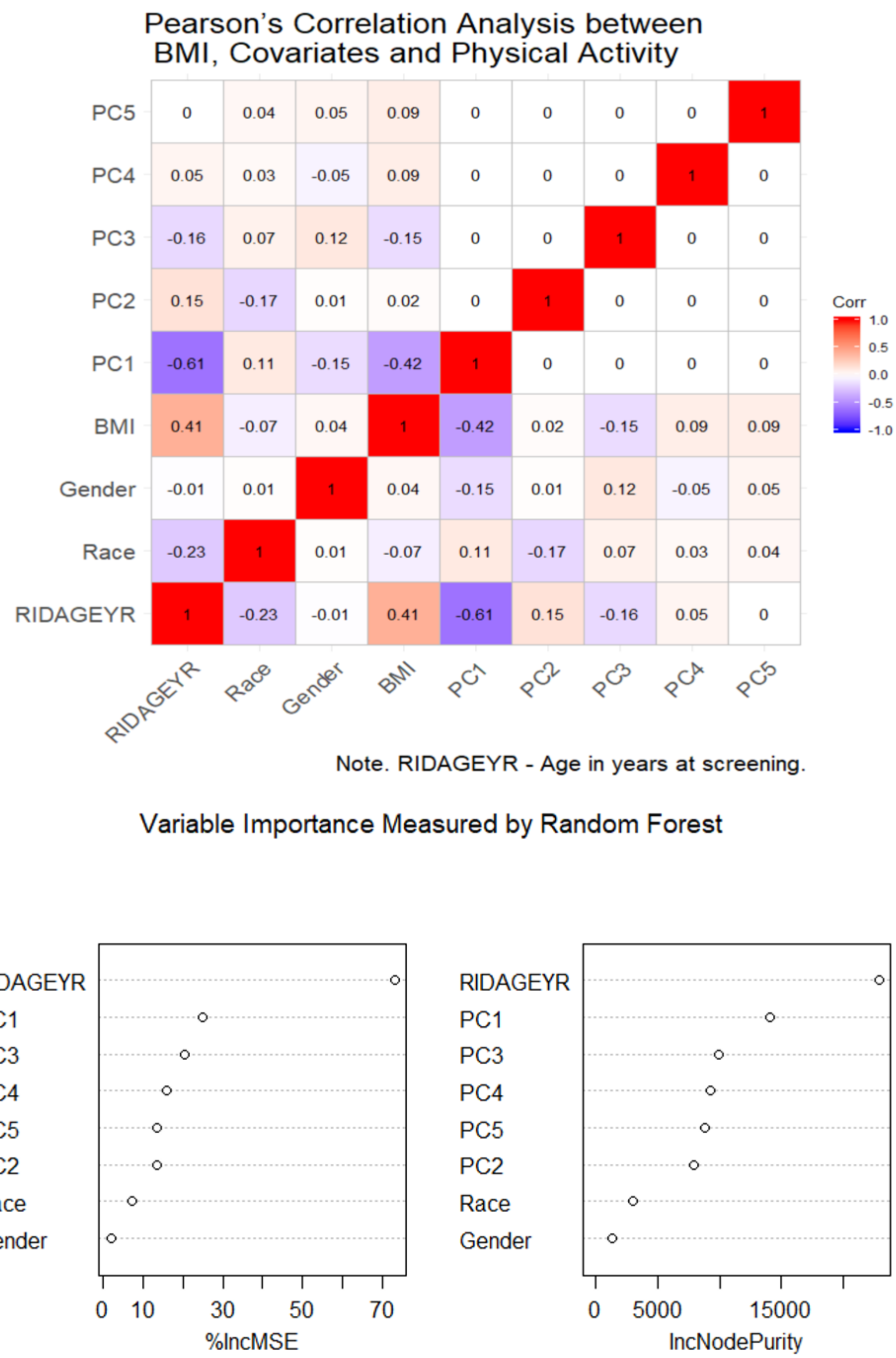

Figure 3. (a) Correlation analysis between BMI, covariates and the first five principal components with Pearson's correlation coefficients. (b) Two measures of variable importance are reported. The former (%IncMSE) is based upon the mean decrease of accuracy in predictions on the out of bag samples when a given variable is excluded from the model. The latter (IncNodePurity) is a measure of the total decrease in node impurity that results from splits over that variable, averaged over all trees.

Model selection

A challenging aspect of PCA is choosing the number of relevant principal components to keep. This can be determined by looking at the cumulative variance explained ratio as a function of the number of components (scree plot). As shown in Figure 2(a), the scree plot of PCA indicates that PC4 and PC5 are inflection points, since the curve flattens out subsequently. Thus, no substantial increase in the cumulative percentage of variance can be explained even adding more PCs. However, inferentially determining the number of PCs remains a difficult task and there is no single approach [17]. To address this problem, 10-fold cross-validation was used [18]. The first 50 PCs were used in cross-validation while always keeping the covariates. The overall root-mean-square error (RMSE) and R-squared would be our metrics when evaluating model performance.

Note that the smallest RMSE or the largest R-squared doesn't necessarily indicate the best number of PCs to use in the model. Overfitting may occur and affect the predictive accuracy of the regression model with a new dataset. For example, according to Figure 4, we may consider using the first 14 PCs in regression models because RMSE and R-squared reach their lowest and highest value at PC 14, respectively. However, one shouldn't include unimportant or irrelevant predictors whose presence adds unnecessary complexity to the model and increases the uncertainty regarding the importance of the remaining predictors. Thus, using the first 14 PCs in the model may lead to an over-specified model that may not reflect reality. As such, t-tests were used to conduct hypothesis tests on the regression coefficients obtained in the linear regression model fits and model selection was performed via significance thresholding. In addition, nested model searchers were performed using F-tests on groupings of variables. Thus, we compared models without any principal components of daily activity and models with different numbers of components ranged from 1 to 14.

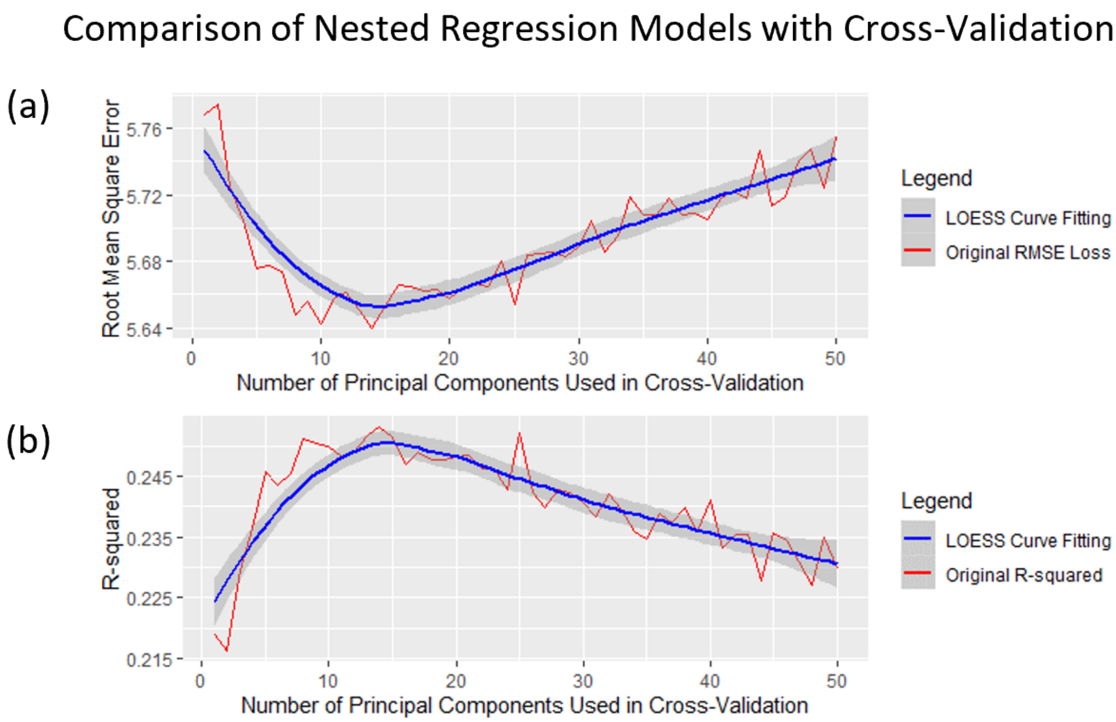

Figure 4. (a) Root-mean-square error computed for nested models with 10-fold cross-validation. The curve reaches its lowest level at the fourteenth principal component. (b) R-squared computed for nested models with 10-fold cross-validation. The curve reaches its highest level at the fourteenth principal component.

Chapter 3

Results

Figure 2(b) suggests that PC1 explains the highest fraction of explained variance (15.97%) and large loadings on PC1 are highly concentrated within daytime hours (6 to 18 hours military time). Thus, the associated PC1 scores primarily represent the person-specific shift in daytime activity. It does, however, have a slight upward trend and a more drastic drop in the evening. PC2 is nonzero in the morning and evening. Therefore, scores associated with this PC represent a compression or increase in the total time spent active. PC3 has opposite signs in the morning and evening and thus the associated scores represent greater or less activity in the morning or evening.

Previous studies have examined age-related differences in the physical activity behaviors of young adults [19]. In figure 3(a), our correlation analysis also indicates the strongest correlation is between PC1 and age (p-value $<$ 2.2e-16). Among all predictors, the variable importance metrics derived from random forest suggest that age affects BMI the most followed by PC scores, age, and sex.

In Figure 4, cross-validation results suggest including the first 14 principal components in the regression model. However, according to Table 1, PCs 8 - 14 do not have robust apparent effects on BMI. In addition, comparing nested models with different numbers of principal components using F-tests showed that models including the first 3, 4, 5, 6, and 8 were significantly preferable to those including also the higher components. Considering these results in aggregate, we focus on the regression model employing the first 8 principal components.

Table 1 also includes summaries of the regression model using age, race, sex, and the first 8 principal components to predict BMI. We tested the null hypothesis that the coefficients were equal to zero (no effect). Predictors with low p-values (<0.05) included age (p-value < 0.001), PA (PC1, p-value < 0.001) and race (Other Hispanic, p-value $= 0.015$). In sum, these results show significant associations between BMI, age, race and physical activity level.

Table 1. Comparison between linear regression models using the first 14 principal components and the first 8 principal components.

	Regression Model using the first 14 PCs		**Regression Model using the first 8 PCs**	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	22.95	**<0.001**	22.94	**<0.001**
RIDAGEYR	0.06	**<0.001**	0.06	**<0.001**
Race [Mexican American]	0.59	0.160	0.65	0.122
Race [Other Hispanic]	0.57	0.604	0.63	0.564
Race [Black]	0.68	0.103	0.67	0.106
Race [Other]	-1.26	0.102	-1.27	0.101
Gender [Female]	0.49	0.123	0.50	0.115
PC1	-0.12	**<0.001**	-0.12	**<0.001**
PC2	-0.00	0.929	-0.00	0.877
PC3	-0.08	**<0.001**	-0.08	**<0.001**
PC4	0.08	**0.001**	0.08	**0.001**
PC5	0.07	**0.006**	0.07	**0.005**
PC6	-0.05	0.071	-0.05	0.062
PC7	-0.02	0.484	-0.02	0.453
PC8	0.11	**0.001**	0.11	**0.001**
PC9			0.03	0.311
PC10			-0.04	0.232
PC11			0.02	0.553
PC12			-0.05	0.251
PC13			-0.03	0.394
PC14			0.03	0.540
Observations	1385		1385	
R2 / R2 adjusted	0.238 / 0.230		0.241 / 0.230	

Note. RIDAGEYR - Age in years at screening.

Chapter 4

Discussion

The primary aim of this study was to investigate the relationship between PA, covariates, and BMI. Here, we have provided a method that helps select the number of PCs in linear regression models. Our results explored associations between PA and BMI among an ethnically diverse sample of adults in the US aged 6 to 85 years. These results, in conjunction with a better understanding of the causes of BMI differences, will have important public health implications for the development of weight gain prevention strategies [20].

A limitation of the regression approach is that we do not consider non-linear associations between predictors and BMI, nor do we consider interactions between predictors. Linear regression is a classical parametric method which requires explicit modeling of non-linearities and interactions, if necessary. Variable importance is not straightforward in linear regression due to correlations between variables. In contrast, our random forest approach is nonparametric and allows non-linearities and variable interactions to be learned from the data without any need to explicitly model them [21]. Our random forest variable importance metrics suggested that activity-based PCs, are more influential on the prediction of BMI than race or sex.

There is also a high percentage of participants with missing BMI data, which may have led to sampling-based bias in model estimates. Moreover, previous research has revealed that PA intentions can vary considerably over time, fluctuating on at least daily, weekly, biweekly, and monthly time scales and that these fluctuations can predict physical activity [22–24]. It remains unclear, however, how these fluctuations in intentions interact with habits to regulate daily physical activity [25]. Finally, our analysis only includes daily and weekly patterns of PA, which might be limited to investigate associations.

References

1. Arroyo-Johnson, C. & Mincey, K. D. Obesity Epidemiology Worldwide. *Gastroenterol Clin North Am* **45,** 571–579 (2016).
2. Hruby, A. & Hu, F. B. The Epidemiology of Obesity: A Big Picture. *Pharmacoeconomics* **33,** 673–89 (2015).
3. Smirnova, E. *et al.* The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: NHANES 2003-2006. *J Gerontol A Biol Sci Med Sci* (2019).
4. Palta, P. *et al.* Self-reported and accelerometer-measured physical activity by body mass index in US Hispanic/Latino adults: HCHS/SOL. *Prev Med Rep* **2,** 824–8 (2015).
5. Ogden, C. L., Carroll, M. D., Kit, B. K. & Flegal, K. M. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA* **311,** 806–14 (2014).
6. Prince, S. A. *et al.* A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int J Behav Nutr Phys Act* **5,** 56 (2008).
7. Troiano, R. P. *et al.* Physical activity in the United States measured by accelerometer. *Med Sci Sports Exerc* **40,** 181–8 (2008).
8. Leroux, A. rnhanesdata: NHANES Accelerometry Data Pipeline (2020).
9. Team, R. C. R: A Language and Environment for Statistical Computing (2019).
10. Watson, K. B. *et al.* The Attributable Proportion of Specific Leisure-Time Physical Activities to Total Leisure Activity Volume Among US Adults, National Health and Nutrition Examination Survey 1999-2006. *J Phys Act Health* **13,** 1192–1201 (2016).
11. Leroux, A. *et al.* Organizing and analyzing the activity data in NHANES. *Stat Biosci* **11,** 262–287 (2019).
12. Choi, L., Liu, Z., Matthews, C. E. & Buchowski, M. S. Validation of accelerometer wear and nonwear time classification algorithm. *Med Sci Sports Exerc* **43,** 357–64 (2011).
13. Hart, T. L., Swartz, A. M., Cashin, S. E. & Strath, S. J. How many days of monitoring predict physical activity and sedentary behaviour in older adults? *Int J Behav Nutr Phys Act* **8,** 62 (2011).
14. Bakrania, K. *et al.* Associations of moderate-to-vigorous-intensity physical activity and body mass index with glycated haemoglobin within the general population: a cross-sectional analysis of the 2008 Health Survey for England. *BMJ Open* **7,** e014456 (2017).

15. Saffer, H., Dave, D., Grossman, M. & Leung, L. A. Racial, Ethnic, and Gender Differences in Physical Activity. *J Hum Cap* **7,** 378–410 (2013).
16. Jackson, A. S. *et al.* The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. *Int J Obes Relat Metab Disord* **26,** 789–96 (2002).
17. Li, Y., Wang, N. & Carroll, R. J. Selecting the Number of Principal Components in Functional Data. *J Am Stat Assoc* **108** (2013).
18. Bro, R., Kjeldahl, K., Smilde, A. K. & Kiers, H. A. Cross-validation of component models: a critical look at current methods. *Anal Bioanal Chem* **390,** 1241–51 (2008).
19. Leslie, E., Fotheringham, M. J., Owen, N. & Bauman, A. Age-related differences in physical activity levels of young adults. *Med Sci Sports Exerc* **33,** 255–8 (2001).
20. Ball, K., Owen, N., Salmon, J., Bauman, A. & Gore, C. J. Associations of physical activity with body weight and fat in men and women. *Int J Obes Relat Metab Disord* **25,** 914–9 (2001).
21. Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician* **63,** 308–319 (2009).
22. Maher, J. P., Doerksen, S. E., Elavsky, S. & Conroy, D. E. Daily satisfaction with life is regulated by both physical activity and sedentary behavior. *J Sport Exerc Psychol* **36,** 166–78 (2014).
23. Conroy, D. E., Maher, J. P., Elavsky, S., Hyde, A. L. & Doerksen, S. E. Sedentary behavior as a daily process regulated by habits and intentions. *Health Psychol* **32,** 1149–57 (2013).
24. Scholz, U., Keller, R. & Perren, S. Predicting behavioral intentions and physical exercise: a test of the health action process approach at the intrapersonal level. *Health Psychol* **28,** 702–8 (2009).
25. Rebar, A. L., Elavsky, S., Maher, J. P., Doerksen, S. E. & Conroy, D. E. Habits predict physical activity on days when intentions are weak. *J Sport Exerc Psychol* **36,** 157–65 (2014).

Curriculum Vitae # Luchao Qi **May, 2020**

3111 North Charles Street
Baltimore, Maryland 21218 USA
(+1) 443.839.9129
lqi9@jhu.edu

EDUCATION AND DEGREES

2018–Present Graduate student, Department of Biomedical Engineering
Johns Hopkins University

2014–2018 Undergraduate student, Department of Biomedical Engineering Northeastern University

RESEARCH EXPERIENCE

Johns Hopkins University, Bloomberg School of Public Health
Department of Biostatistics *(05 2019-Present)*
Graduate Student in the laboratory of Brian S. Caffo

- Explored associations between BMI and physical activity through regression analysis.
- Developed R package MRIPCA for principal component analysis on T1 volumetric analysis of MRIcloud output.

TECHNIQUES AND SKILLS

Computer Science Techniques

Machine Learning ; Deep Learning ; Bioinformatics ; Biostatistics

Computing and Other Skills

- Basic user of Linux/Unix systems
- Perl programming
- R programming
- Python programming

PUBLICATIONS

Non-contact High-frequency Ultrasound Microbeam Stimulation: A Novel Finding and Potential Causes of Cell Responses. Luchao Qi, Qi Zhang, Yan Tan, Kwok-ho Lam, Hairong Zheng, Ming Qian. *IEEE Transactions on Biomedical Engineering* **67**. PAGES (1074 - 1082).

Calcium fluorescence response of human breast cancer cells by 50-MHz ultrasound microbeam stimulation. Luchao Qi, Qi Zhang, Kwok Ho Lam, Ruibiao Guo, Runkang Chen, Jiangbang Huang, Ruoyan Meng, Zhaoyang Wang, Hairong Zheng, Ming Qian. *2017 IEEE International Ultrasonics Symposium (IUS)*.